Volume 17 Number 19 1989 Nucleic Acids Research

Polymers of random short oligonucleotides detect polymorphic loci in the human genome

Gilles Vergnaud

Laboratoire de Génétique Moléculaire, Centre d'Etudes du Bouchet, BP no. 3, 91710 Vert-le-Petit, France

Received July 18, 1989; Revised and Accepted September 5, 1989

ABSTRACT

Polymers of random 14 mer oligonucleotides are shown to detect discrete loci in the human genome. Eighteen different synthetic tandem repeats of random 14 base-pair units (STRs) have been generated and all of them turn out to detect polymorphic loci on southern blots of human DNA samples, presumably corresponding to a variable number of tandem repeats (VNTR). This finding suggests that minisatellites are a major component of the human genome and are strongly associated with the generation of genetic variability. In addition, it should open new strategies to make new polymorphic probes available.

INTRODUCTION

Redundancies are present at different scales in the human genome: satellite structures spanning thousands of kilobases (for a review see 1); 'minisatellites' (2, 3) also called VNTRs (4) (for Variable Number of Tandem Repeats) when they are associated with RFLPs (for Restriction Fragment Length Polymorphisms), spanning a few kilobases; and very short redundancies spanning a few nucleotides, as described by Tautz et al.(5).

On an evolutionary point of view, minisatellites are usually considered as selfish DNA or byproducts of the genome metabolism. Computer simulations of the evolution of sequences submitted to recombination showed the progressive generation of tandem repeats (6). The same general phenomenon might underlay satellite and minisatellite formation, as well as overrepresentation of simple motifs. The work I will present here is based on two main hypotheses. The first one is that minisatellites are much more frequent than usually thought. The second one is that the sequence of the basic unit is not always the key factor of the amplification process.

If this is true, synthetic tandem repeats (STRs) made of short random units could be able to detect RFLPs when probed on, for example, Southern blots of human DNA. In a previous work (7), I have shown that STRs made from simple (as defined in 4) sequence units could detect polymorphic loci. As a further test of these hypotheses, the work presented here reports the results obtained using eighteen different STRs made now from random 14-mer units.

MATERIALS AND METHODS

Cloning of the polymers

The oligonucleotides were purchased from the Pasteur Institute, Paris. They were purified on a denaturing acrylamide gel, then phosphorylated, annealed and ligated using standard procedures (7,8). The T4 polynucleotide kinase and ligase were purchased from Boehringer. Briefly, 2 μg from both complementary oligonucleotides are phosphorylated in a 50 μl

PROBE	SEQUENCE	% G+C	WASHING	CRITERION
14C1	GATCGCTCTCTCGA	57%	4xSSC 60°C	27°C
14C2	GGCAGGATTGAAGC	57%	2xSSC 60°C	22°C
14C3	AGCTAAGCCTAGCA	50%	1xSSC 55°C	28°C
14C4	AACTGCCCCCTCTT	57%	3xSSC 63°C	24°C
14C5	GACAAACAGAGCAA	43%	1xSSC 55°C	25°C
14C6	CGAGCCAAACGCTA	57%	1xSSC 50°C	36°C
14C7	TCCTAGAATTTTCT	29%	3xSSC 45°C	30°C
14C8	GGCCGTAGCGCGGT	79%	1xSSC 50°C	45°C
14C9	ATGCACCGTCGCAC	64%	1xSSC 60°C	28°C
14C10	GTGAAAACGTCTTCC	50%	3xSSC 45°C	38°C
14C11	AGTCATGGTAGAGC	50%	1xSSC 55°C	27°C
14C12	GTTTCTCCAACAGA	43%	1xSSC 50°C	30°C
14C13	AGCCGTCTGTTTTC	50%	1xSSC 50°C	33°C
14C14	CTGAAAACGATGGG	50%	1xSSC 55°C	28°C
14C15	CCGTAGCAGGTAGA	57%	1xSSC 50°C	36°C
14C16	GGTAGAGGCAACTC	57%	3xSSC 57°C	30°C
14C17	CAAAAGTCAGGCGT	50%	1xSSC 55°C	28°C
14C18	TGATTTAAGTCCAA	29%	3xSSC 45°C	30°C
14C19	GGGTGCTCGGGTAC	71%	1xSSC 60°C	31°C
14C20	CCCCGCTCAGGTAC	71%	3xSSC 55°C	37°C
16C1	AACAGCTATGACCATG	44%	3xSSC 57°C	24°C

Table 1: Sequence of the polymerised units.
The table lists the 5′−3′ sequences of the monomers with their G+C content, the washing conditions used to obtain the results shown in figure 1 and 2 and an estimation of the corresponding criterion. The melting temperature (Tm) of the tandem repeats has been estimated from Tm = 81.5 + 0.41(%G+C) + 16.6 log(Na+)−(300 + 2000 (Na+))/L, (10) where L is the length of the DNA fragment and is approximately 100 bp after random primed labelling. The criterion is the difference between the Tm and the washing temperature.

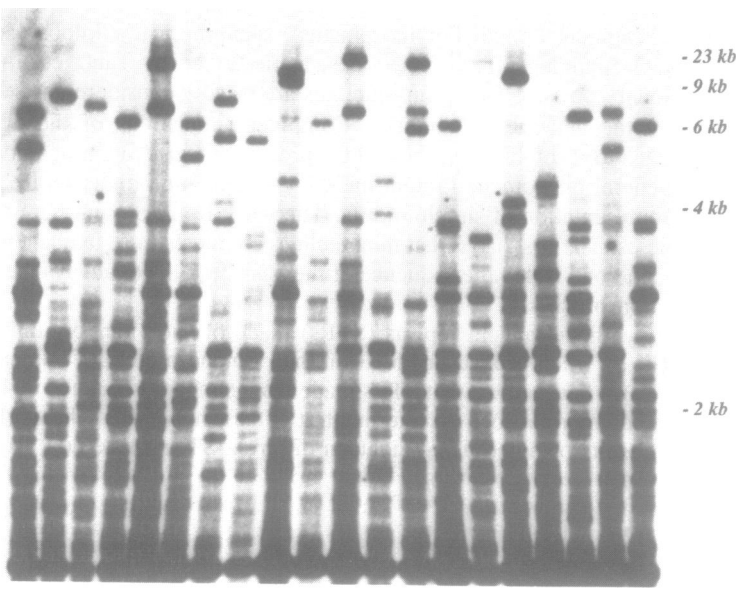


Figure 1: STR 14C2 hybridised on DNA samples from 20 unrelated individuals.
The DNA samples were digested with Hinf I. The filter was washed at 60°C in 2×SSC, 0.1% SDS. The size scale in kilobases is given on the right.

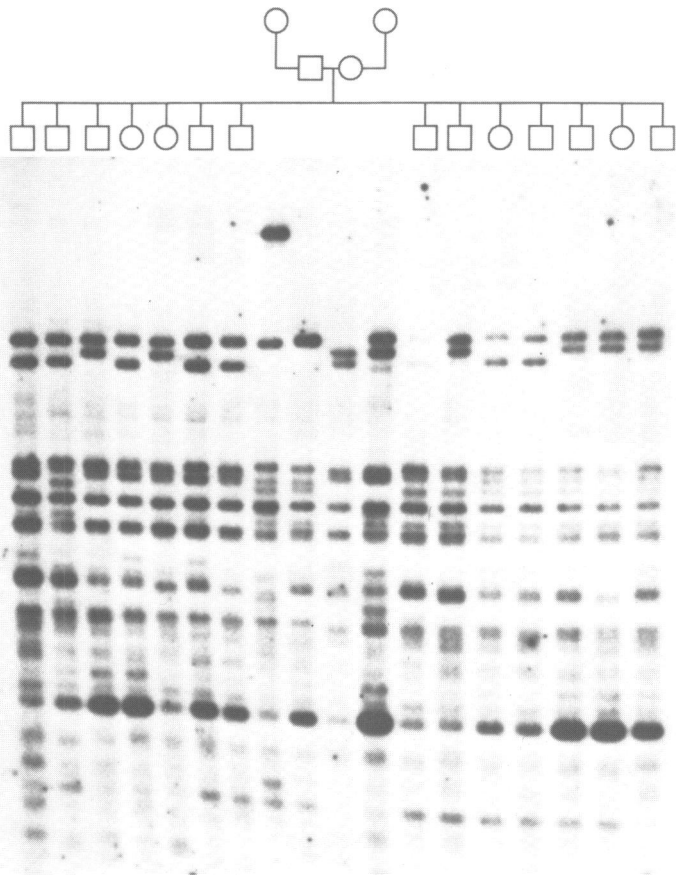
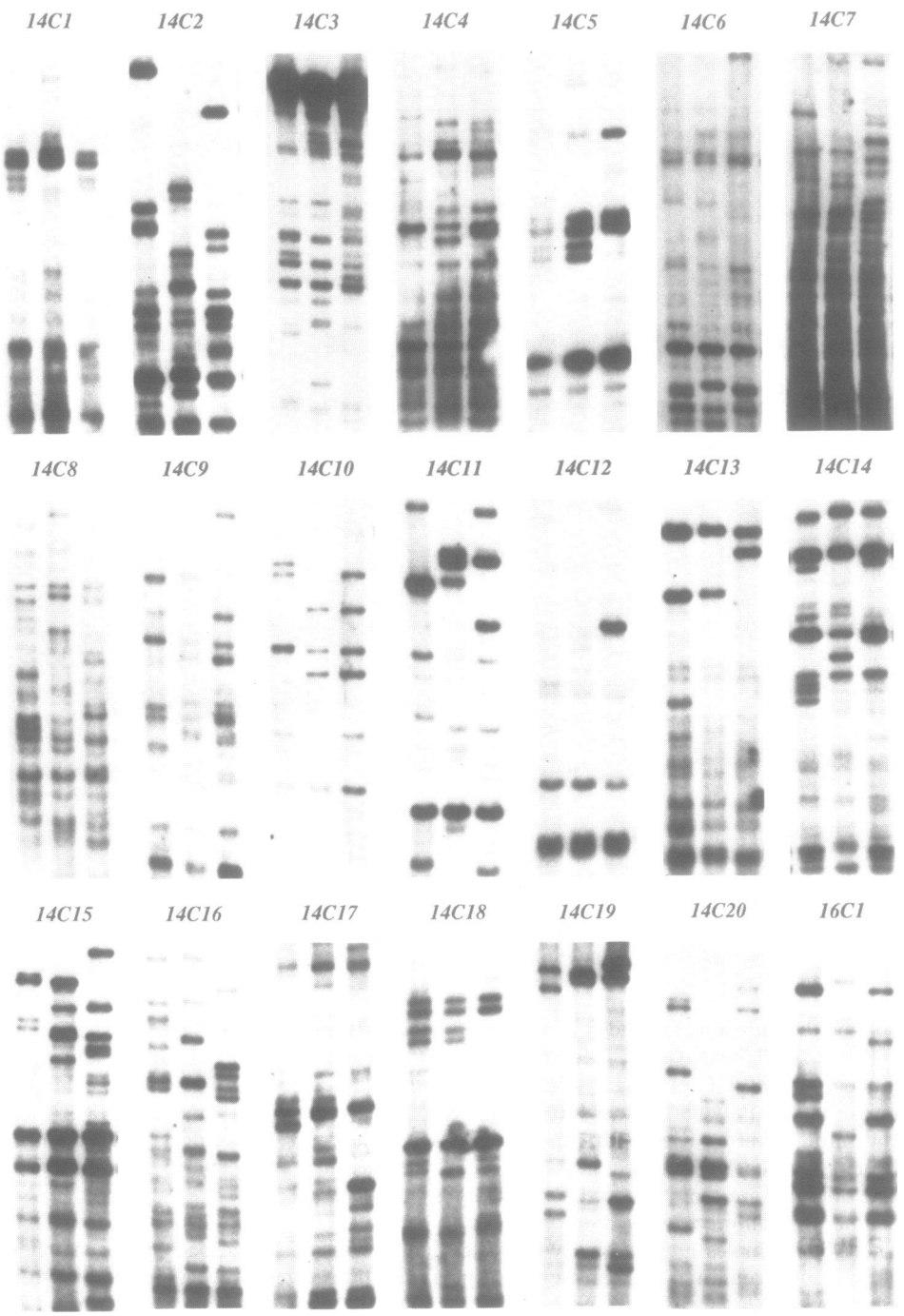


Figure 2: STR 14C2 hybridised on DNA samples from a family. The DNA samples were digested with Hinf I and the filter was washed at 60°C in 2×SSC, 0.1% SDS. Males are indicated by squares, and females by circles.

volume then precipitated and resuspended in 3 μl of 10 mM Tris pH 7.5, 1mM EDTA, 100 mM NaCl, heated at 65°C for 15 minutes and left at 4°C for a few hours. The ligation is then done at 4°C overnight after dilution with 50 μl of 1× ligation mixture and one unit of ligase. Polymers above 400 bp are then purified from an acrylamide gel (8). After end-filling with the Klenow fragment of DNA polymerase I (Boehringer), an aliquot is ligated in a Puc 18 SmaI vector (Appligene) and cloned in a TG1 or DH5 alpha host (9). The colonies are then screened for the largest insert (10).

Southern blotting and hybridisations

The DNA samples are digested with restriction enzymes as described by the supplier (Appligene), run on a 1% agarose gel (Sigma) and transferred by vacuum blotting to a Genescreen nylon membrane (NEN). The vacuum transfer unit was purchased from LKB and is used as recommended except that the length of each step is increased: 15 minutes


14C1
14C2
14C3
14C4
14C5
14C6
14C7
14C8
14C9
14C10
14C11
14C12
14C13
14C14
14C15
14C16
14C17
14C18
14C19
14C20
16C1

with HCl 0.25M; 40 minutes with NaOH 0.5M-NaCl 1.5M; 30 minutes with Tris-NaCl; 30 minutes with 20×SSC. The DNA is then fixed by UV irradiation. The filters are prehybridised (15 minutes or more) and hybridised (overnight) at 50°C in 2% SDS, 0.45 M Na_2PO_4 pH 7.2, 0.5% skimmed milk, 1 mM EDTA (derived from 11) with random primed DNA insert (12) at a probe concentration of approximately 10^6 dpm/ml. The washings are done as indicated in table 1 with 0.1% SDS for at least one hour. The filters are autoradiographed at −80°C for 1 to 4 days with two amplifying screens and a XAR 5 Kodak film.

RESULTS

In this set of experiments, eighteen STRs were synthesised from random 14-mer oligonucleotides. The length chosen for the basic units is an arbitrary choice. Their sequence was designed at random using a computer. In order to easily generate STRs from each of the units, the 18 complementary 14-mer oligonucleotides were also purchased so that annealed together two complementary motifs would form tandem arrays a few hundred bases long (7). These arrays were cloned with insert sizes ranging from 200 to 600 base pairs. The inserts were used as probes on different Southern blots containing restriction digests of human DNA samples. Table 1 gives the sequence of each motif, the washing conditions used and an estimation of the corresponding criterion (13). Three additional STRs are also presented: STRs 14C19 and 14C20 are derived from 14 base-pair long adaptors we use in the laboratory for unrelated purposes. They differ with each other only at the first four nucleotides and at the ninth, however their hybridisation profiles are very different. STR 16C1 is derived from the 16 bases M13 reverse sequencing primer. Figure 1 shows the result obtained with one of the cloned STRs (14C2) hybridised on DNA from 20 unrelated individuals. Bands are detected in a wide range of sizes. On this panel, all the individuals appear to give a different pattern. Different polymorphic loci seem to be detected since different washing conditions can give slightly different patterns with variations in the relative intensities of the bands (data not shown). Fig. 2 shows the same probe hybridised on a three generation family blot (the grandfathers are missing). As can be expected for a multilocus probe, the patterns are less variable since the parents will be homozygotes for some of the loci. The segregation of the upper bands appears to be compatible with a Mendelian inheritance of one polymorphic locus. The two grandmothers are heterozygotes, as is the mother, but the father is homozygote. Each child receives one allele from his or her mother. Some other polymorphisms can also be detected in the lower part of the blot. Similar work was done using all the STRs generated and Figure 3 shows the results obtained on three unrelated individuals (the three individuals shown are not always the same ones). A relevant fact is that all the STRs do detect a discrete number of loci, some of which are polymorphic. However, some of the probes do not detect many different polymorphisms and will give sometimes identical patterns among some of the twenty individuals tested (14C1, 14C4, 14C5, 14C18) when the others are similar to 14C2 and reveal different images for each of the twenty unrelated individuals tested (data not shown). In the case of 14C18, this fact could be linked to the very low

Figure 3: Hybridisation of the probes with unrelated individuals.
The DNA samples were digested with the restriction enzyme Hae III (14C7, C8, C11, C12, C13, C16, C20) or HinfI. The washing conditions were done as indicated in Table 1.

GC content. However, 14C1 and 14C4 have GC contents identical to other probes such as 14C2 which seems to be detecting many highly variable loci (Fig. 1). 14C10 and 14C11 also have identical GC contents. 14C11 detects different highly polymorphic loci at an estimated criterion of 30°C, but 14C10 detects no cross-hybridising locus when used with that same criterion of 30°C (data not shown). The question whether these are probe specific behaviours independent of the species tested and reflecting a more general feature of genome evolution and DNA sequences generation is presently being investigated. However, the present results already suggest at least some sequence specificity that remains to be explored.

DISCUSSION

At this stage, I can only speculate on the nature of the sequences detected with these STR probes. They most probably are tandem repeats similar to the structures used as probes since the stringency used for hybridisation and washing should not allow the detection of unique oligomers. In addition, they detect RFLPs independently of the restriction endonuclease used (HaeIII, HinfI, EcoRI, HindIII; data not shown) suggesting that the polymorphisms detected are of the insertion-deletion type. Further studies are under way to test this hypothesis and provide an estimation of the number of minisatellites in the genome.

Human gene mapping using Restriction Fragment Length Polymorphisms (RFLPs) and linkage analysis is most efficient when highly informative polymorphic loci can be identified. In the past few years, probes detecting such loci have been isolated from genomic DNA, usually by chance (11; for review see 15, 16), or through the use of a screening strategy involving previously caracterised highly polymorphic probes (2, 4, 17, 18). These probes were often shown to be minisatellites, that is tandem repeats of short subunits, polymorphic because of variations in the number of the tandem repeats. Recently, the polymerase chain reaction (PCR) has been proposed as a way to analyse much smaller size variations within microsatellites (19, 20) which could make available a larger number of polymorphic loci. The present work supports the hope that STRs will provide a new approach enabling the cloning of multiallelic systems unrelated to previously described families, if they exist. At the present stage, STRs used directly provide a very wide range of new multilocus probes since many other random oligonucleotides, of 14 base-pairs and other lengths, can be used. No function has been clearly assigned to these tandemly-repeated structures. I believe that the genome has in this way the possibility of experimenting new sequences that in a minority of cases will eventually turn out to be of use. Some genes such as, among others, the period gene of Drosophila (21), the DMD gene (22) or some of the genes encoding proteins containing zinc finger domains (23), contain tandem repeats. However, it is difficult in some cases to know whether this kind of gene structure is present from the beginning or is the result of progressive duplications of an ancestral gene. In that respect, the involucrine gene (24, 25, 26) is noteworthy. An important part of its coding region is found in primates only. It is mainly composed of approximately forty repeats of a 30 base-pair unit. Ohno has suggested that tandem arrays could be at the origin of some ancestral (27) or even modern (28) genes, because tandem arrays of short motifs will very often constitute open reading frames (29). If the creation of tandem repeats at the DNA level is a significant process for the creation of new genes, then one would expect tandem repeats of all kinds to be frequent components of the genome since only a minority of them will eventually turn out to be selectively advantageous. Some regions of the human genome such as the subtelomeric area do appear to be very rich in minisatellite structures

since Jeffreys and coll. have described subtelomeric cosmid clones containing at least two unrelated such structures (30, 31).

The quest for redundancies in DNA sequences is usually done using databases.The approach described here can be seen as a search for redundancies of the minisatellite type in an ultimate database since it is performed on the whole genome. For that reason it is not subjected to the biases present in databases that contain mostly transcribed DNA sequences of often very 'old' genes. As a result, this work shows for example that an observation such as the detection of hypervariable loci using a M13-derived tandem repeat (32) is part of a much wider phenomenon: any polymer of a short unit can lead to a similar observation.

Starting from the working hypotheses that the initiation of minisatellite formation is essentially sequence independent (33) and is a widespread phenomenon, I have demonstrated that it is possible to detect polymorphic loci with synthetic polymers of short random sequences. However the data presented here shows important variations from one probe to the other regarding the number and the degree of polymorphisms detected. Clearly, this indicates at least some degree of sequence specificity in the amplification process. It will be interesting to see if there are recombinogenic 'cores' in the genome apart from the one described by Jeffreys and coll. The sequences I generated and used to detect polymorphic loci are not significantly related to any sequences in the Genbank database (release 55.0). Irrespective of the nature of the DNA that hybridises to these synthetic sequences, the polymorphisms they detect might make possible the identification of new highly polymorphic loci, task undertaken until now through random screenings or through the search of sequence similarities with previously described consensus minisatellite motifs.

ACKNOWLEDGEMENTS

I thank Valérie Lauthier and Monique Zoroastro for technical assistance. I thank Christine Pourcel for her constant help and support throughout this work. I thank Jairam Lingappa for his encouragements; Hanh Nguyen, Fotis C. Kafatos and Gabriel A. Dover for helpful discussions; the members of U.R.E.G. at the Pasteur Institute, Paris, for their help. Jean Weissenbach and Claude Gazin suggested important improvements on my last draft. The DNA samples are provided by the C.E.P.H.(Centre d'Etude du Polymorphisme Humain).

REFERENCES:

1. Beridze, T. (1986) Satellite DNA. Springer, Berlin Heidelberg New York.
2. Jeffreys, A.J., Wilson, V. & Thein, S.L. (1985) Nature **314**, 67–73.
3. Wyman, A. R. and White, R. (1980) Proc. Natl. Acad. Sci. USA **77**, 6754–6758.
4. Nakamura, Y., Leppert, M., O'Connel, P., Wolff, R., Holm, T., Culver, M., Martin, C., Fujimoto, E., Hoff, M., Kumlin, E., and White, R. (1987) Science **235**, 1616–1622.
5. Tautz, D., Trick, M. & Dover, G. (1986) Nature **322**, 652–656.
6. Smith, G. P. (1973) Cold Spring Harbor Symp. Quant. Biol. **38**, 507–513.
7. Vergnaud, G. (1988) Ph. D. Thesis, University of Paris VII.
8. Maniatis, T., Fritsch, E. F. and Sambrook, J. (1982) Molecular Cloning: a laboratory manual. Cold Spring Harbor Laboratory, Cold Spring Harbor.
9. Chung, C. T. and Miller, R. H. (1988) Nucleic Acids Res. **16**, 3580.
10. Del Sal, G., Manfioletti, G. and Schneider, C. (1988) Nucleic Acids Res. **16**, 9878.
11. Church, G. M. and Gilbert, W. (1984) Proc. natn. Acad. Sci. U.S.A. **81**, 1991–1995.
12. Feinberg, A.P. and Vogelstein, B. (1983) Anal. Biochem. **132**, 6–13.
13. Meinkoth, J. and Wahl, G. (1984) Anal. Biochem. **138**, 267–284 .
14. Schum, J. W., Knowlton, R. G., Braman, J. C., Barker, D. F., Botstein, D., Akots, G., Brown, V., Gravius,

T., Helms, C., Hsiao, K., Rediker, K., Thurston, J. and Donnis-Keller, H. (1988) Am. J. Hum. Genet. **42**, 143–159.
15. Pearson, P. L., Kidd, K. K. and Willard, H. F. (1987) Cytogenet. Cell Genet. **46**, 390–566.
16. Kidd, K. K. et al., (1988) Cytogenet. Cell Genet. **49**, 132–218.
17. Wong, Z., Wilson, V., Jeffreys, A. J., and Thien, S. L. (1986) Nucleic Acids Res. **14**, 4605–4616.
18. Nakamura, Y., Carlson, M., Krapcho, K., Kanamori, M. and White, R. (1988) Am. J. Hum. Genet. **43**, 854–859.
19. Weber, J. L. and Paula, E. M. (1989) Am. J. Hum. Genet. **44**, 388–396.
20. Litt, M. and Luty, J. A. (1989) Am. J. Hum. Genet. **44**, 397–401.
21. Yu, Q. et al., (1987) Nature **326**, 765–769.
22. Koenig, M., Monaco, A. P., and Kunkel, L. M. (1988) Cell **53**, 219–228.
23. Miller, J., Mc Lachlan, A. D. and Klug, A. (1985) EMBO J. **4**, 1609–1614.
24. Eckert, R.L. & Green, H. (1986) Cell **46**, 583–589.
25. Tseng, H. and Green, H. (1988) Cell **54**, 583–589.
26. Teumer, J. and Green, H. (1989) Proc. natn. Acad. Sci. U.S.A. **86**, 1283–1286.
27. Ohno, S. (1984) J. molec. Evol. **20**, 313–321.
28. Ohno, S. (1984) Proc. natn. Acad. Sci. U.S.A. **81**, 2421–2425.
29. Ohno, S. and Jabara, M. (1986) Chemica Scripta **26B**, 43–49.
30. Royle, N. J., Clarkson, R. E., Wong, Z. and Jeffreys, A. J. (1988) Genomics 3, 352–360.
31. Armour, J. A. L., Wong, Z., Wilson, V., Royle, N. J. and Jeffreys, A. J. (1989) Nucleic Acids Res. **17**, 4925–4935.
32. Vassart, G., Georges, M., Monsieur, R., Brocas, H., Lequarre, A. S. and Christophe, D. (1987) Science **235**, 683–684.
33. Dover, G.A. (1980) Nature **285**, 618–620.

This article, submitted on disc, has been automatically converted into this typeset format by the publisher.